APPLIED PHYSICS

Rapid discovery of stable materials by coordinate-free coarse graining

Rhys E. A. Goodall[1], Abhijith S. Parackal[2], Felix A. Faber[1], Rickard Armiento[2]*, Alpha A. Lee[1]*

A fundamental challenge in materials science pertains to elucidating the relationship between stoichiometry, stability, structure, and property. Recent advances have shown that machine learning can be used to learn such relationships, allowing the stability and functional properties of materials to be accurately predicted. However, most of these approaches use atomic coordinates as input and are thus bottlenecked by crystal structure identification when investigating previously unidentified materials. Our approach solves this bottleneck by coarse-graining the infinite search space of atomic coordinates into a combinatorially enumerable search space. The key idea is to use Wyckoff representations, coordinate-free sets of symmetry-related positions in a crystal, as the input to a machine learning model. Our model demonstrates exceptionally high precision in finding unknown theoretically stable materials, identifying 1569 materials that lie below the known convex hull of previously calculated materials from just 5675 ab initio calculations. Our approach opens up fundamental advances in computational materials discovery.

INTRODUCTION

Finding a needle in a haystack is often used as an analogy for materials discovery. Only a small proportion of viable material compositions [believed to be of the order $\mathcal{O}(10^{10})$ (*1*)] will have thermodynamically stable polymorphs that are experimentally accessible. Most approaches to tackling this challenge focus on predictive models for materials properties, metaphorical sieves that filter out the hay. Here, we seek an alternative approach: Can we cut down the size of materials space by changing how we represent materials—making most of the hay disappear?

Our approach is motivated by a concept ubiquitous in science: coarse graining. Taking molecular chemistry, for example, chemists typically build intuitions about chemical properties using molecular graphs. Molecular graphs are a coarse-grained representation of molecules, with each graph corresponding to a unique ensemble of atomic coordinates. Searching in the enumerable space of molecular graphs, as opposed to the innumerable space of possible atomic coordinates, has enabled the development of powerful computational tools (*2*, *3*) and efforts that exhaustively enumerate chemical space (*4*, *5*).

In materials science, however, an analogous coarse-grained representation of crystal structures is missing. Thus, we are left confronting the innumerable search space problem. Composition-based approaches can somewhat overcome this challenge (*6*–*9*) but do so at the cost of discarding all information about the crystal structures of the materials being considered. Hence, either extensive computational crystal structure searching or laboratory-based experiments are required to validate predictions.

One avenue to maneuver around this challenge has been to explore restricted classes of structure prototypes using novel descriptors, e.g., perovskites (*10*–*12*), quaternary Heuslers (*13*), or elpasolites (*14*). Specifying the prototype avoids the need for crystal structure searching, empowering more extensive screening campaigns as the computational cost of validation is greatly reduced.

Here, we introduce an approach that generalizes these prototype-restricted models by considering Wyckoff representations, coordinate-free sets of symmetry-related positions in a crystal. This framework allows us to develop accurate machine learning models for materials discovery tasks where the relaxed crystal structure is a priori unknown.

We first test the ability of our model to identify previously unidentified stable materials across a diverse range of chemistries, showing that it has a precision ~3 times larger than that of state-of-the-art coordinate-free methods based on elemental substitutions (*15*, *16*). We then evaluate the performance of our model in identifying stable structures within phase diagrams with diverse structures, showing that our model finds low-energy structures in the phase diagram with ~5 times lower computational effort. Last, we develop a materials exploration pipeline that, starting from an initial nucleus of known materials, screens nearby materials space and allows the efficient discovery of new stable materials. We identify 1569 hitherto unknown materials that are below the known convex hull of previously calculated materials from just 5675 ab initio calculations.

RESULTS

Wyckoff representation regression

Building an accurate machine learning model hinges on identifying model inputs that are sufficiently informative to allow the target variable to be predicted. However, for a machine learning model to be useful in practice, these inputs need to be significantly cheaper to obtain than the cost of labeling data. In the context of materials discovery, previous works have shown that virtual screening workflows based on Kohn-Sham density functional theory (DFT) can be used to identify novel functional materials (*17*, *18*). Separately, it has been shown that accurate machine learning models can be built for the formation energies of inorganic crystals calculated via DFT using the DFT-relaxed crystal structure as the model input (*19*–*22*). Inference using these models is significantly cheaper than the DFT calculations they approximate, but sadly, their application to materials discovery is circular because arriving at a DFT-relaxed structure necessitates calculating the energy using DFT multiple times.

[1]Department of Physics, University of Cambridge, Cambridge, UK. [2]Department of Physics, Chemistry and Biology, Linköping University, Linköping, Sweden.
*Corresponding author. Email: rickard.armiento@liu.se (R.A.); aal44@cam.ac.uk (A.A.L.)

Several groups have therefore proposed to use composition-based inputs (*6*–*9*), which avoid the upfront need for structure identification. However, the composition is not expressive enough to differentiate polymorphs. This is a significant shortcoming as different polymorphs can have radically different properties, most famously the example of diamond and graphite. Hence, we turn to model inputs that can distinguish polymorphs while also avoiding the cost of DFT. Such models can be used to triage which DFT calculations are carried out in materials discovery workflows, allowing for a more efficient use of computational resources.

In crystallography, one way to completely specify the crystal structure of a material is via a combination of (i) the space group of the structure, (ii) the dimensions of its unit cell, and (iii) a set of Wyckoff positions with the elements that sit on them. The Wyckoff positions describe sites that map onto equivalent sites under the symmetry transformations of the given space group (*23*). As a consequence, a single Wyckoff position can encode the positions of multiple atoms. To construct model inputs from sets of Wyckoff positions, we discard the information about the exact positions and lattice parameters. In the resulting coordinate-free representation, the Wyckoff representation, each Wyckoff position is simply labeled by a Wyckoff letter and the element at that position. Consequently, as the Wyckoff representation is discrete, it is possible to computationally enumerate Wyckoff representations that represent candidate materials for use in screening campaigns.

The procedure of obtaining the Wyckoff representation from a crystal structure can be viewed as a coarse-graining process that takes us from an unsymmetrized initial parameter space of size $4N + 6$, through the symmetrized Wyckoff position space of maximum size $5M + 6$, to the much smaller coordinate-free space of Wyckoff representations with size $2M$, where N is the number of sites in the unit cell, and the corresponding number of Wyckoff positions M satisfies $M \leq N$. The back mapping from the coarse-grained Wyckoff representation to the full structure can often be satisfactorily obtained via a single symmetry-constrained DFT relaxation of a prototype structure (see Fig. 1).

To use the Wyckoff representation as the input for a machine learning model, we formulate the task of property prediction as a multiset regression problem. A message passing neural network architecture based on the Roost architecture (*8*) is used to do this; the Roost model performs materials property prediction via set regression on the weighted set of elements in a material's composition.

The principal idea behind the model architecture is to embed the coordinate-free Wyckoff positions of a given material into a vector space. The representations in this embedded space are then updated via message passing operations that consider all directed pairwise combinations of members in the multiset. The messages propagate contextual information between Wyckoff positions, leading to the emergence of material-specific representations. These message passing stages are repeated multiple times before a permutation invariant pooling operation is applied to the multiset to get a fixed-length representation. As the labeling of Wyckoff positions includes several choices of setting, we carry out on-the-fly augmentation of equivalent Wyckoff representations. We then average the fixed-length representations for these equivalent inputs to ensure invariance to this choice. These averaged fixed-length representations are then fed into a feed-forward output neural network that returns the model predictions.

This work focuses primarily on models that predict the formation energy of inorganic crystalline materials, although the proposed framework and inputs are applicable to any material property. We call the proposed model Wren (**W**yckoff **Re**presentation regressio**N**). Throughout this work, we train deep ensembles consisting of 10 Wren models starting from different random initializations (*24*), allowing us to estimate the model's uncertainty and providing better point estimates. Details of the Wren architecture and the hyperparameters used are given in the Supplementary Materials.

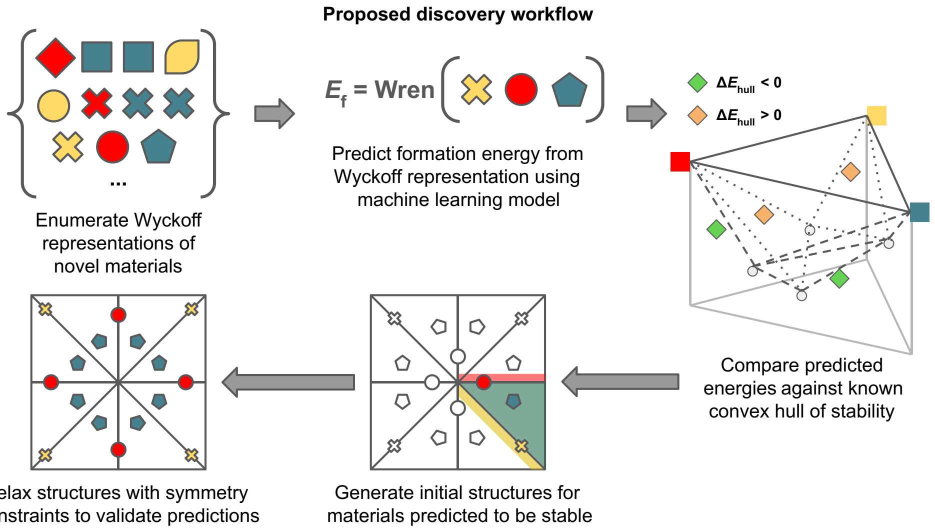
Proposed discovery workflow
...
Enumerate Wyckoff representations of novel materials
E_f = Wren
Predict formation energy from Wyckoff representation using machine learning model
$\Delta E_{hull} < 0$
$\Delta E_{hull} > 0$
Compare predicted energies against known convex hull of stability
Generate initial structures for materials predicted to be stable
Relax structures with symmetry constraints to validate predictions

Fig. 1. Coarse-graining materials space using Wyckoff representations enables efficient data-driven materials discovery. A machine learning–powered materials discovery workflow that takes advantage of the benefits of the proposed Wyckoff representation. The workflow uses a machine learning model to predict formation energies for candidate materials in an enumerated library of Wyckoff representations (shapes are used to denote different Wyckoff positions and colors to denote different element types). These predicted formation energies are then compared against the known convex hull of stability. Structures satisfying the required symmetries are then generated and relaxed for materials predicted to be stable. The calculated energies of the relaxed structures can then be compared against the known convex hull to confirm whether the candidate is stable.

Selecting stable materials from diverse chemical space

To accelerate the screening of materials space for novel stable materials, a model must reduce the expected number of calculations needed to find a candidate below the known convex hull [here taken to be the convex hull of the Materials Project (MP) dataset before cleaning]. We first assess the ability of the model to generalize across materials space to unseen combinations of elements.

To do this, we consider two datasets. (i) The MP database (*25*) is a highly curated database of high-throughput DFT calculations. At the time of access, the MP database contains approximately 140,000 crystal structures. We apply a canonicalization and cleaning treatment (see Materials and Methods) that leaves a final MP dataset containing approximately 105,000 distinct materials. (ii) The dataset of Wang, Botti, and Marques (WBM), obtained from (*16*), contains calculated energies and properties of a large number of crystal structures that were generated through the substitution of elements in known crystal structures from MP with chemically similar elements (*26*). Hence, the WBM dataset chemically extrapolates from the MP dataset. After deduplication and cleaning, the WBM dataset contains approximately 215,000 materials.

We make predictions for the formation energies of the materials contained in the WBM dataset using a Wren model trained on the MP dataset. We then assess how well the Wren model selects potentially stable materials from the WBM dataset. The relevant metrics are the following: the prevalence, the proportion of materials below the known convex hull (actual positives); the precision, how many of the predictions of potentially stable materials are correct (i.e., the ratio of true predicted positives to the total predicted positives); and the recall, how many of the actual materials below the known convex hull are found (i.e., the ratio of true predicted positives to actual positives). In this setup, the ratio of the precision and the prevalence gives the enrichment factor or degree of acceleration. Enrichment factors are frequently reported for virtual screening campaigns in drug discovery applications (*27*, *28*).

The precision using the Wren model to triage calculations is 38%. Consequently, given that the prevalence of theoretically stable materials in the WBM dataset is 15%, using Wren leads to an enrichment factor of 2.5. As enrichment here is computed with respect to the active search strategy of (*16*), this translates into a significant improvement in efficiency over random or exhaustive search strategies as our improvements compound multiplicatively with theirs. Consequently, triaging screening workflows based on Wren should enable more materials below the known hull to be identified with limited computational resources. We also observe a high recall of 76%, meaning that Wren misses relatively few potentially stable materials.

The screening performance of the model can be tuned by adjusting our triage criteria. For example, an alternative triage criterion would be to require that $\Delta\widehat{E}_{\text{Hull-Pred}} + \hat{\sigma} < 0$, where $\Delta\widehat{E}_{\text{Hull-Pred}}$ is the predicted distance of a candidate material from the known convex hull and $\hat{\sigma}$ is the predictive uncertainty of the model. This uncertainty-adjusted criterion encourages the model to suggest candidates that it is more certain about, leading to an increased precision of 53%. The enrichment factor for the uncertainty-adjusted criterion is 3.5. Consequently, the choice of triage criteria should depend on the aims of a given workflow: the expected opportunity cost of false negatives versus false positives, the availability of experimental or computational resources, and how easy it is to expand the candidate pool.

The strong performance of Wren can be explained by looking at how the mean absolute error changes as a function of the distance from the known convex hull. Figure 2 shows that near to the stability threshold, $\Delta E_{\text{Hull-MP}} = 0$, Wren makes highly accurate predictions of the formation energy.

More significant errors are seen for materials far above and far below the hull. However, in these regions, the average error is less than the energy to the convex hull, meaning that the model's classifications are still reliable. The large errors far above the hull are due to the routine underestimation of the formation energy of unstable structures. This underestimation is a manifestation of a bias in the MP dataset toward structures with low formation energies. The bias arises from the fact that large numbers of the initial structures in the MP dataset are sourced from the Inorganic Crystal Structure Database (ICSD) (*29*). This result highlights the importance of negative examples for building generally applicable machine learning models (*30*–*32*).

Exploration of unseen tertiary phase diagrams

From an applications perspective, researchers are often interested in exploring a single or small number of chemical systems that have not previously been studied. Typical approaches for computationally mapping out the convex hull of novel chemical systems (*33*, *34*) are highly expensive, often requiring thousands of structures to be relaxed.

To evaluate the ability of Wren to assist when mapping the phase diagrams of targeted chemical systems, we consider the dataset of Tholander, Andersson, Armiento, Tasnadi, and Alling (TAATA) (*35*), consisting of three highly sampled ab initio phase diagrams for the Hf-Zn-N, Ti-Zn-N, and Zr-Zn-N ternary systems. The ternary systems studied in the TAATA dataset were investigated for their potential in piezoelectric devices and energy harvesting applications.

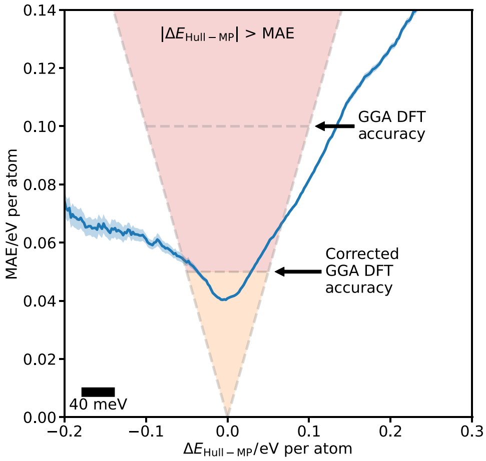

Fig. 2. Wren's average error is below DFT error in the region around the stability threshold. Rolling mean absolute error (MAE) on the WBM dataset as the energy to the convex hull is varied for Wren model. A scale bar is shown for the windowing period of 40 meV per atom used when calculating the rolling average. The SEM is shaded around each curve. The highlighted V-shaped region shows the area in which the average absolute error is greater than the energy to the known convex hull; this is the region where the model is most at risk of misclassifying structures. In most of this region, Wren's accuracy is well below the threshold of 100 meV per atom considered to be the accuracy of semilocal DFT across diverse chemistries (*66*) and comparable to the threshold of ~50 meV per atom characteristic of fitted correction schemes (*67*–*69*).

We focus on ternary systems due to the fact that while crystal structure prediction approaches such as those in (*33*, *34*) work very well for unary and binary systems, there has been less work applying these methods to ternary systems (*36*–*38*) because of the combinatorial explosion in the number of candidates that need to be relaxed per phase diagram to obtain reliable results—often in excess of 10,000 relaxations need to be carried out for each chemical system.

The TAATA dataset contains a diverse range of stable and unstable structures for each composition [full details about the construction of the TAATA dataset are given in the Supplementary Materials and in (*35*)]. After applying a canonicalization and cleaning treatment (see Materials and Methods), we are left with 3104 entries over 523 compositions in the Ti-Zn-N phase diagram, 2711 entries over 453 compositions in the Zr-Zn-N phase diagram, and 3381 entries over 596 compositions in the Hf-Zn-N phase diagram.

We trained Wren on the MP dataset but excluded all the tertiary compounds from the Ti-Zn-N, Zr-Zn-N, and Hf-Zn-N chemical systems. This model was then used to predict the energies of tertiary compounds in the TAATA dataset. These formation energy predictions were then used to construct hypothetical convex hulls for the Ti-Zn-N, Zr-Zn-N, and Hf-Zn-N chemical systems (see the Supplementary Materials). Figure 3 shows how selecting which relaxations to carry out on the basis of the predicted distances to the hypothetical convex hulls constructed using the Wren model's predictions can accelerate phase diagram exploration. To quantify this effect, we look at the enrichment factor as a function of the number of calculations. The enrichment factor describes the ratio between the number of candidates found satisfying a target criterion when using a given virtual screening strategy and the number of positive candidates that hypothetically would have been found if the candidates were screened randomly. Considering materials within 20 meV per atom of the DFT-calculated convex hull as our target criteria, we see that after 250 calculations, we have a high enrichment factor of 5.4 in the Ti-Zn-N chemical system, 5.1 for the Zr-Zn-N chemical system, and 4.5 for the Hf-Zn-N chemical system when using the Wren model, i.e., we are saving 4.5 to 5.4 times the computational resource compared to a random search.

Computational prospecting for previously unidentified stable materials

Having established the promise of Wren in predicting the stability of unseen materials, we deploy Wren on the prospective challenge of finding unknown theoretically stable materials. For this stage, we trained Wren on the union of the MP and WBM datasets. This combined dataset contains approximately 322,000 materials after canonicalization and cleaning. We randomly sampled 5% of the dataset to use as a test set and trained on the remaining 95%. The resulting model has a mean absolute error of 31 meV per atom on this test set, which is below the commonly quoted chemical accuracy level of 1 kcal/mol (43 meV per atom) (*39*). The model's accuracy as a function of training set size follows a power-law relationship (see the Supplementary Materials). Reassuringly, the model does not appear to saturate in performance, suggesting that the representation is rich enough, and further increases in model performance can be unlocked given more data (*40*–*42*).

While the coarse-grained space of Wyckoff representations is computably enumerable and far smaller than the infinite space of atomic coordinates, attempting materials discovery by exhaustively screening all possible Wyckoff representations for general chemistries is computationally inefficient as the prevalence of stable materials remains vanishingly low even in the coarse-grained space. To effectively tackle the task of materials discovery in general chemistries, it is necessary to first construct a design space with a higher expected prevalence (*13*). To do this, we draw inspiration from previous work (*15*, *16*) and generate candidates for screening by making elemental substitutions in crystal structures that are near to the known convex hull. Focusing on the use of machine learning to accelerate workflows tackling general chemistries is particularly compelling as false-positive and false-negative data produced in such workflows can subsequently be used to retrain the model. Inclusion of a diverse range of negative examples in the training data is key to improving performance in targeted exhaustive workflows, such as in Section C, where design space enrichment is not applicable.

To obtain our substitution probabilities, we extracted 39,164 ordered structures from the ICSD (*29*, *43*) and binned them according to their Wyckoff representations. Within each prototype, all pairs of

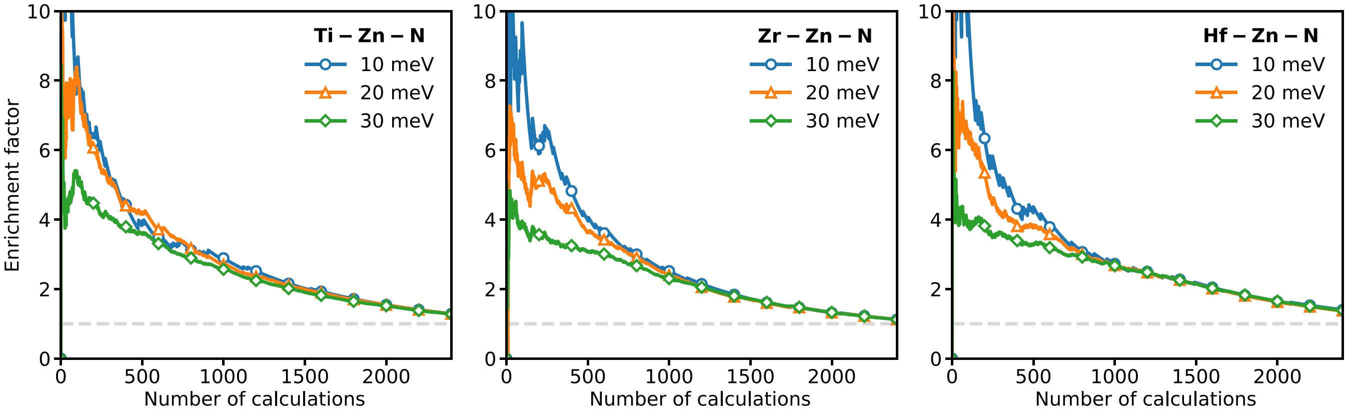

Fig. 3. Wren accelerates the recovery of low-energy structures in unseen chemical systems. The figures show how the enrichment factor varies as we use Wren to direct the exploration of the Ti-Zn-N, Zr-Zn-N, and Hf-Zn-N chemical systems. The enrichment factor is the ratio of candidates found satisfying a given triage criterion to the number we would expect to find via a random search. The enrichment factor is plotted for candidates within 10, 20, and 30 meV per atom from the convex hull of the full explored system. A light-gray guideline is included to show the performance expected from a random model, an enrichment factor of 1. The plots demonstrate that using Wren leads to a significant degree of early enrichment of low-energy structures.

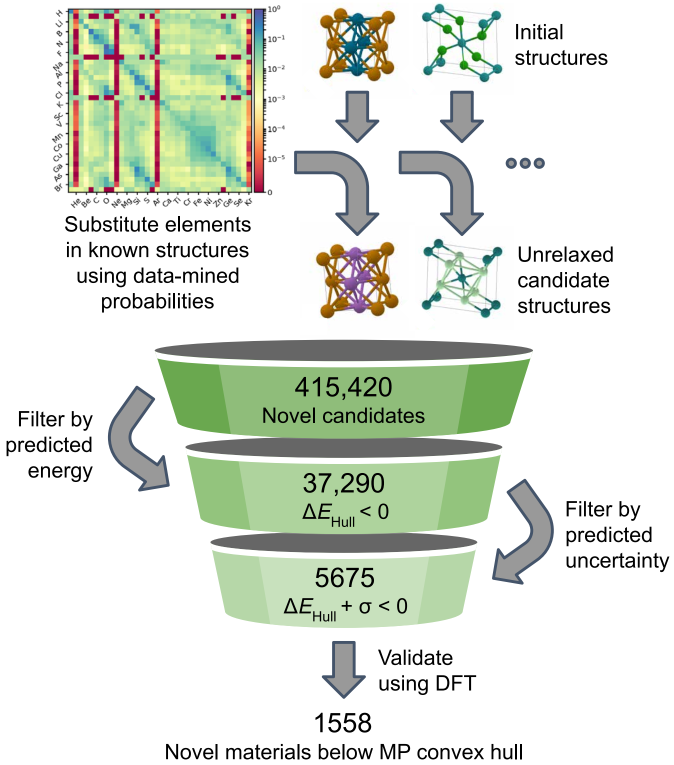
Initial structures
Substitute elements in known structures using data-mined probabilities
Unrelaxed candidate structures
415,420
Novel candidates
Filter by predicted energy
37,290
$\Delta E_{\text{Hull}} < 0$
Filter by predicted uncertainty
5675
$\Delta E_{\text{Hull}} + \sigma < 0$
Validate using DFT
1558
Novel materials below MP convex hull

Fig. 4. Wren enables automated computational prospecting of previously unidentified stable materials. Data-mined substitution probabilities are used to generate candidates for screening. A heatmap of the data-mined log substitution probabilities for the first 36 main group elements is shown in the top left. The matrix captures known chemical trends, for example, that halogens can often be substituted for each other in crystal structures. Using the Wren allows far more unrelaxed candidates to be considered than possible in conventional DFT-led high-throughput workflows. The funnel diagram shows the number of unrelaxed candidates that pass the different stability criteria when filtering based on the predictions of the Wren model. In total, 4721 of 5675 validation calculations completed. Of these, 1569 were below the known convex hull, giving a precision of 33% among the completed calculations.

structures are compared, and we count which element substitutions (including self-substitutions) would be needed to change one structure into the other (*26*). We only consider substitutions where all Wyckoff positions sharing one element type are changed simultaneously and not per position substitutions. Once normalized, the rows of the count matrix can be interpreted as substitution probabilities for each element.

Using these data-mined probabilities, we generated a screening library of materials by substituting different elements into structures taken from the MP dataset. We only consider initial structures from the MP dataset with energies above the convex hull less than 100 meV per atom. This choice of this threshold means that we should be including most metastable structures within the MP dataset. We consider 10 different substitutions for each initial structure. Candidates that have the same composition as materials already present in the union of the MP and WBM datasets are removed from the library. Lanthanide- and actinide-based materials and materials containing noble elements were also excluded. This workflow produced a screening library of approximately 415,000 candidates. The size of the screening library can be readily increased by considering more elemental substitutions per structure.

Despite constraining our screening set to be close to known materials, it is likely that we are still asking the model to make predictions in areas of materials space where it lacks support from the training data. As shown on the WBM dataset, uncertainty estimation allows us to reduce the risk in our materials screening process by factoring our model's uncertainty into our triage criterion. For simplicity, we use the same simple uncertainty-adjusted criterion considered previously; $\Delta\widehat{E}_{\text{Hull-Pred}} + \hat{\sigma} < 0$. In total, 5675 candidates satisfied this screening criterion (see Fig. 4).

Validation with DFT resulted in 4721 completed calculations across 4464 unique compositions. Of these, 1569 structures were confirmed to be below the convex hull of the MP dataset and 1369 below the convex hull of the union of the MP and WBM datasets. Therefore, the precision for the completed calculations was 33% with respect to the MP convex hull, confirming the workflow's ability to accelerate materials discovery. Although direct comparisons are not strictly permissible, as previous prospective searches using machine learning have been restricted to single prototypes, the Wyckoff representation–based approach presented significantly surpasses previously reported precisions of 4% (*14*) and 13% (*22*). Another key consideration for materials discovery is whether the model is able to generalize and make novel discoveries or simply interpolate current knowledge. Of the 4721 completed calculations, 269 were assigned to canonicalized prototypes for which there were no isopointal prototypes in the training set [see materials and methods for description of canonicalization and cleaning approach]. Of these, 78 were confirmed to be below the convex hull of the union of the MP and WBM datasets. Developing workflows to directly target the discovery of structures for which no isopointal prototypes exist remains a key challenge for future work.

DISCUSSION

In this work, we introduced the concept of using coarse graining to accelerate materials discovery. We developed the framework of Wyckoff representation regression, Wren, and applied it to predict the formation energy of materials. Wren collapses an infinite search space of atomic coordinates into a combinatorially enumerable search space, enabling efficient data-driven exploration of materials space. On a set of challenging tasks curated from the literature, we find that our approach can accurately map the phase diagrams of unseen chemical systems and is ~3 times better at finding stable materials than current methods based on elemental substitutions.

We developed a materials prospecting pipeline using Wren. As a proof of concept, we identified 1558 new materials below the known convex hull from just 5675 calculations. These results demonstrate that leveraging Wren allows for more efficient and extensive expansion of computational material science databases. Such efforts are crucial to expedite the search for a wide variety of industrially desirable materials required for the transition to a low-carbon economy, e.g., thermoelectrics (*44*), piezoelectrics (*35*), fast-ion conductors (*45*), high-voltage multivalent cathode materials (*46*), and caloric materials (*47*).

MATERIALS AND METHODS

Wren model architecture

The bulk of the Wren architecture directly mimics that of Roost (*8*), and we refer the readers there for an in-depth description of how the message passing is formulated. The principle difference between

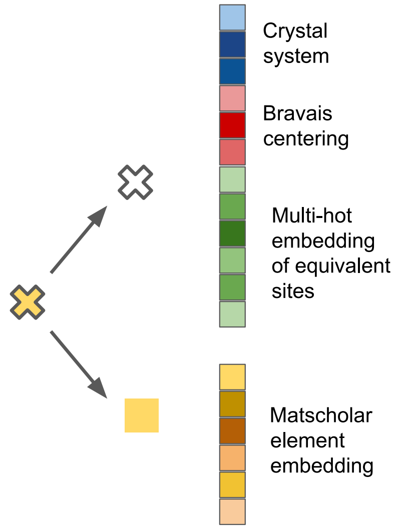
Crystal system
Bravais centering
Multi-hot embedding of equivalent sites
Matscholar element embedding

Fig. 5. Breakdown of different components of the Wyckoff position embeddings. The Wyckoff position embeddings are made up of two parts: first, the Wyckoff proportion of the embedding that is composed of three subsections encoding the crystal system, Bravais centering, and equivalent sites in the Wyckoff positions; second, the elemental embedding for which we take the matscholar embedding from (*48*).

the two architectures comes in that the nodes on the dense graph now represent the different Wyckoff positions rather than the different elemental species (see Fig. 5). The elemental information is encoded using the "matscholar" embedding from (*48*), which has a dimensionality of $d_{\mathrm{el}} = 200$. The remainder of the node embedding comprises the Wyckoff position embedding (described below) plus the fractional multiplicity of that Wyckoff position within the unit cell. The combined dimensionality of the Wyckoff proportion of the embedding is $d_{\mathrm{wyk}} = 445$.

To reduce the total dimension of the node embeddings, we project both the elemental and Wyckoff embeddings into lower-dimensional spaces using learnt affine transformations. The low-dimensional embeddings are then concatenated to give the node embeddings. In this work, we chose $d^*_{\mathrm{el}} = 32$ and $d^*_{\mathrm{wyk}} = 32$, giving a total dimensionality of $d = 64$ for the node embeddings.

We use three message passing layers, each with a single attention head. We chose single–hidden layer neural networks with 256 hidden units and LeakyReLU activation functions for both parts of the soft-attention mechanism. The output network consists of a feed-forward neural network with skip connections and ReLU activation functions. The output network used has four hidden layers containing 256, 256, 128, and 64 hidden units, respectively.

Wyckoff position embedding

In total, across the 230 crystallographic space groups in 3D, there are 1731 different Wyckoff positions. The embedding we use is made up of three parts: a one-hot encoding of the crystal system (of which there are six), a one-hot encoding of the Bravais lattice centering (of which there are five), and an encoding constructed from the sum of multi-hot encodings of the equivalent sites within a given Wyckoff position (see Fig. 5). To construct the multi-hot encodings, we first collate all the sites within all the allowed Wyckoff positions as recorded on the Bilbao crystallographic server (*49*). Each site can be broken apart into its offset and algebraic terms (whether the position corresponds to a line, a plane, etc.), e.g.

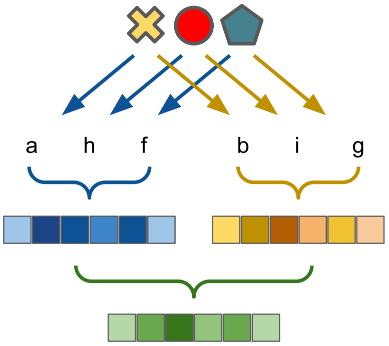

Fig. 6. On-the-fly augmentation of equivalent Wyckoff representations ensures invariance to equivalent descriptions. The labeling of Wyckoff positions includes a choice of setting; to ensure that our model is invariant to these choices, we perform on-the-fly augmentation of all equivalent Wyckoff representations and then average the augmented embeddings before they are fed into the output network.

$$(-x + y + 1/4, y, z + 3/4) = (1/4, 0, 3/4) + (-x + y, y, z) \tag{1}$$

From this, we construct separate one-hot encodings for the unique algebraic and unique offset positions. We end up with 185 unique algebraic positions and 248 unique offset positions. A Wyckoff position is then represented by a sum of the embeddings of all the allowed sites. The resulting embedding has a dimensionality of 444 with the 1731 Wyckoff positions being encoded into 1038 unique embeddings. This embedding is designed to try and expose as many possible correlations as possible that might exist between different Wyckoff positions. As an illustrative example, the embeddings for the f Wyckoff position of *Fm*3 (no. 202), *F*432 (no. 209), and *Fm*3*m* (no. 225) are all the same. This arises as they are all face-centered cubic lattices that describe the positions of 32 atoms within the unit cell at [(0, 0, 0), (0, 1/2, 1/2), (1/2, 0, 1/2), (1/2, 1/2, 0)] ⊕ [(*x*, *x*, *x*), (−*x*, −*x*, *x*), (−*x*, *x*, −*x*), (*x*, −*x*, −*x*), (−*x*, −*x*, −*x*), (*x*, *x*, −*x*), (*x*, −*x*, *x*), (−*x*, *x*, *x*)], where *x* is a free coordinate of the Wyckoff positions. The embedding vector has 4's in the positions corresponding to the (*x*, *x*, *x*), (−*x*, −*x*, *x*), (−*x*, *x*, −*x*), (*x*, −*x*, −*x*), (−*x*, −*x*, −*x*), (*x*, *x*, −*x*), (*x*, −*x*, *x*), (−*x*, *x*, *x*) algebraic terms and 8's in the positions corresponding to the (0, 0, 0), (0, 1/2, 1/2), (1/2, 0, 1/2), (1/2, 1/2, 0) offset terms. In principle, further engineering of this embedding could be carried out to encode more prior knowledge; however, for the sizes of dataset considered in this work, the benefit of doing so is likely to be marginal.

Invariance to equivalent Wyckoff representations

The categorization of Wyckoff positions depends on a choice of origin (*50*). Hence, there is not a unique mapping between the crystal structure and the Wyckoff representation. To ensure that the model is invariant to the choice of origin, we perform on-the-fly augmentation of Wyckoff positions with respect to this choice of origin (see Fig. 6). The augmented representations are averaged at the end of the message passing stage to provide a single representation of equivalent Wyckoff representations to the output network. By pooling at this point, we ensure that the model is invariant and that its training is not biased toward materials for which many equivalent Wyckoff representations exist.

Evaluation of Wyckoff positions

For this work, we primarily make use of spglib (*51*) to determine the space group and Wyckoff positions for the structures in the datasets. We set the tolerance thresholds as 0.1 Å for positions and 5° for angles (note that these are the same tolerances as used in MP to calculate the space group). In real materials, we often observe some degree of off-site relaxation away from high-symmetry sites. Depending on the level of anisotropy and the symmetry finder's tolerance threshold, this might result in materials being classed as *P*1. As lower-symmetry Wyckoff representations encode less information about the structure, the symmetry finder tolerance is an important hyperparameter to bear in mind. However, preliminary investigations showed that varying the tolerance threshold between typical values of 0.01 and 0.1 Å did not significantly affect model accuracy on the TAATA dataset. We note that as an alternative to manual selection of tolerance hyperparameters, symmetry finders with adaptive tolerances, such as `aflow-sym` (*52*), could be used for the identification of the space group and Wyckoff positions. However, given that we did not observe any appreciable improvement in accuracy using `aflow-sym` and adaptive schemes are typically associated with greater computational cost, `spglib` was picked over other symmetry finders because of its speed.

Model training

Throughout this work, we train deep ensembles of 10 models starting from different random initializations for each data setup and architecture considered. All the models examined in this work were trained using the AdamW optimizer (*53*) with a fixed learning rate of 3 × 10^{-4}. A mini-batch size of 128 and a weight decay parameter of 10^{-6} were used for all the experiments. The models were trained for 400 epochs (cycles through the training set).

Formally, deep ensembles require the use of a proper scoring rule for training. Therefore, we train all models to minimize the following robust L1 loss function, which is an example of a proper scoring rule for regression (*54*, *55*)

$$\mathcal{L} = \sum_i \frac{\sqrt{2}}{\hat{\sigma}_{a,\theta}(x_i)} \| y_i - \hat{\mu}_\theta(x_i) \|_1 + \log(\hat{\sigma}_{a,\theta}(x_i)) \qquad (2)$$

where $\hat{\mu}_\theta(x_i)$ and $\hat{\sigma}_{a,\theta}(x_i)^2$ are a predictive mean and predictive aleatoric variance outputted by the model and y_i is the target label.

The expectations, $\hat{y}(x_i)$, and epistemic uncertainties, $\hat{\sigma}_a(x_i)^2$ from the ensemble are calculated as

$$\hat{y}(x_i) = \frac{1}{W} \sum_w^W \hat{\mu}_{\theta_w}(x_i) \qquad (3)$$

$$\hat{\sigma}_e^2(x_i) = \frac{1}{W} \sum_w^W (\hat{y}(x_i) - \hat{\mu}_{\theta_w}(x_i))^2 \qquad (4)$$

where the index *w* runs over the *W* members of the ensemble. The total uncertainty of the ensemble expectation is simply the sum of the epistemic contribution and the average of the aleatoric contributions from each model in the ensemble

$$\hat{\sigma}^2(x_i) = \hat{\sigma}_e^2(x_i) + \frac{1}{W} \sum_w^W \hat{\sigma}_{a,\theta_w}^2(x_i) \qquad (5)$$

Canonicalization and cleaning

All the data used to train models in this work went through a canonicalization and cleaning process. Tables 1 to 3 show how much data are discarded at each stage.

The canonicalization stage removes higher-energy structures that have equivalent Wyckoff representations to other structures in the dataset. We adopt the same canonicalization scheme as used by the AFLOW prototype encyclopedia (*56*, *57*). Most structures removed by canonicalization are triclinic, as the lack of symmetries in triclinic systems results in many distinct structures mapping to the same Wyckoff representation.

For the WBM dataset, we carried out cleaning based on the relaxed structures and relaxed Wyckoff representations. We removed elemental structures in the WBM dataset to ensure that our end points for calculating formation energies were consistent. The union of the MP and WBM datasets used to train the Wren model for the

Table 1. Table showing the impact of cleaning criteria on the MP dataset.

Filter	Number
Full dataset	139,367
Lowest-energy canonical representations	129,190
Formation energy less than 5 eV per atom	129,176
Less than 16 Wyckoff positions	108,656
Less than 64 sites in crystal structure	105,057
Volume per site less than 500 Å$^{\circ3}$	104,878

Table 2. Table showing the impact of cleaning criteria on the WBM dataset.

Filter	Number
Full dataset	257,486
Lowest-energy canonical representations	224,498
After removal of duplicates found in MP	217,085
Excluding pure systems	216,877
Formation energy less than 5 eV per atom	216,859
Less than 16 Wyckoff positions	216,819
Less than 64 sites in crystal structure	216,807
Volume per site less than 500 Å$^{\circ3}$	216,806

Table 3. Table showing the impact of cleaning criteria on the TAATA dataset.

Filter	Number
Full dataset	12,815
Lowest-energy canonical representations	9688
Less than 16 Wyckoff positions	9490
Less than 64 sites in crystal structure	9190
Volume per site less than 500 Å$^{\circ3}$	9190

prospective validation made use of an earlier canonicalization scheme, which led to it containing some duplicated Wyckoff representations. In total, the union of MP and WBM used contained 322,915 materials.

Prospective DFT settings

The validation of predictions in our materials prospecting pipeline was carried out using Kohn-Sham DFT with the plane-wave pseudopotential code VASP (*58*, *59*). Projector augmented wave–type pseudo-potentials (*60*, *61*) were used with the Perdew-Burke-Ernzerhof generalized gradient approximation (GGA) exchange correlation functional (*62*). All calculations were done using a 520-eV plane-wave energy cutoff. The pseudo-potentials and Hubbard *U* values were selected to ensure compatibility with data contained in the MP. The MP MaterialsProjectCompatibility correction scheme implemented in pymatgen was applied to allow the mixing of GGA and GGA + *U* calculations (*63*). We used the High-Throughput Toolkit (httk v1.0) introduced in (*64*) to manage the calculations.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abn4117

REFERENCES AND NOTES

1. D. W. Davies, K. T. Butler, A. J. Jackson, A. Morris, J. M. Frost, J. M. Skelton, A. Walsh, Computational screening of all stoichiometric inorganic materials. *Chem* **1**, 617–627 (2016).
2. D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R. P. Adams, Convolutional networks on graphs for learning molecular fingerprints, in *Proceedings of Advances In Neural Information Processing Systems 28* (Curran Associates, Inc., 2015), pp. 2224–2232.
3. J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).
4. L. Ruddigkeit, R. Van Deursen, L. C. Blum, J.-L. Reymond, Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **52**, 2864–2875 (2012).
5. J.-L. Reymond, The chemical space project. *Acc. Chem. Res.* **48**, 722–730 (2015).
6. B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, C. Wolverton, Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **89**, 094104 (2014).
7. L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, 16028 (2016).
8. R. E. A. Goodall, A. A. Lee, Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *Nat. Commun.* **11**, 1–9 (2020).
9. A. Y.-T. Wang, S. K. Kauwe, R. J. Murdock, T. D. Sparks, Compositionally restricted attention-based network for materials property predictions. *npj Comput. Mater.* **7**, 1–10 (2021).
10. J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, M. A. L. Marques, Predicting the thermodynamic stability of solids combining density functional theory and machine learning. *Chem. Mater.* **29**, 5090–5103 (2017).
11. H. Liu, J. Cheng, H. Dong, J. Feng, B. Pang, Z. Tian, S. Ma, F. Xia, C. Zhang, L. Dong, Screening stable and metastable ABO_3 perovskites using machine learning and the materials project. *Comput. Mater. Sci.* **177**, 109614 (2020).
12. A. Jain, T. Bligaard, Atomic-position independent descriptor for machine learning of material properties. *Phys. Rev. B* **98**, 214112 (2018).
13. K. Kim, L. Ward, J. He, A. Krishna, A. Agrawal, C. Wolverton, Machine-learning-accelerated high-throughput materials screening: Discovery of novel quaternary heusler compounds. *Phys. Rev. Mater.* **2**, 123801 (2018).
14. F. A. Faber, A. Lindmaa, O. A. Von Lilienfeld, R. Armiento, Machine learning energies of 2 million elpasolite (ABC_2D_6) crystals. *Phys. Rev. Lett.* **117**, 135502 (2016).
15. G. Hautier, C. Fischer, V. Ehrlacher, A. Jain, G. Ceder, Data mined ionic substitutions for the discovery of new compounds. *Inorg. Chem.* **50**, 656–663 (2011).
16. H.-C. Wang, S. Botti, M. A. L. Marques, Predicting stable crystalline compounds using chemical similarity. *npj Comput. Mater.* **7**, 1–9 (2021).
17. G. Ceder, Y.-M. Chiang, D. R. Sadoway, M. K. Aydinol, Y.-I. Jang, B. Huang, Identification of cathode materials for lithium batteries guided by first-principles calculations. *Nature* **392**, 694–696 (1998).
18. A. Jain, Y. Shin, K. A. Persson, Computational predictions of energy materials using density functional theory. *Nat. Rev. Mater.* **1**, 1–13 (2016).
19. K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, K.-R. Müller, SchNet—A deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).
20. T. Xie, J. C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
21. C. Chen, W. Ye, Y. Zuo, C. Zheng, S. P. Ong, Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **31**, 3564–3572 (2019).
22. C. W. Park, C. Wolverton, Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Phys. Rev. Mater.* **4**, 063801 (2020).
23. R. W. G. Wyckoff, *The Analytical Expression Of The Results Of The Theory Of Space-groups*, vol. 318. (Carnegie Institution Of Washington, 1922).
24. B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, in *Proceedings of Advances In Neural Information Processing Systems 30* (Curran Associates, Inc., 2017), pp. 6402–6413.
25. A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials* **1**, 011002 (2013).
26. H. Glawe, A. Sanna, E. K. U. Gross, M. A. L. Marques, The optimal one dimensional periodic table: A modified pettifor chemical scale from data mining. *New J. Phys.* **18**, 093011 (2016).
27. A. Bender, R. C. Glen, A discussion of measures of enrichment in virtual screening: Comparing the information content of descriptors with increasing levels of sophistication. *J. Chem. Inf. Model.* **45**, 1369–1375 (2005).
28. N. Huang, B. K. Shoichet, J. J. Irwin, Benchmarking sets for molecular docking. *J. Med. Chem.* **49**, 6789–6801 (2006).
29. G. Bergerhoff, I. D. Brown in *Crystallographic Databases* (International Union of Crystallography, 1987), pp. 77–95.
30. P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, A. J. Norquist, Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
31. J. Schmidt, M. R. G. Marques, S. Botti, M. A. L. Marques, Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **5**, 1–36 (2019).
32. M. K. Horton, S. Dwaraknath, K. A. Persson, Promises and perils of computational materials databases. *Nat. Comput. Sci.* **1**, 3–5 (2021).
33. C. J. Pickard, R. J. Needs, Ab initio random structure searching. *J. Phys. Condens. Matter* **23**, 053201 (2011).
34. Y. Wang, J. Lv, L. Zhu, Y. Ma, CALYPSO: A method for crystal structure prediction. *Comput. Phys. Commun.* **183**, 2063–2070 (2012).
35. C. Tholander, C. B. A. Andersson, R. Armiento, F. Tasnadi, B. Alling, Strong piezoelectric response in stable $TiZnN_2$, $ZrZnN_2$, and $HfZnN_2$ found by *ab initio* high-throughput approach. *J. Appl. Phys.* **120**, 225102 (2016).
36. A. G. Kvashnin, C. Tantardini, H. A. Zakaryan, Y. A. Kvashnina, A. R. Oganov, Computational search for new W-Mo-B compounds. *Chem. Mater.* **32**, 7028–7035 (2020).
37. Z. Lu, B. Zhu, B. W. B. Shires, D. O. Scanlon, C. J. Pickard, Ab initio random structure searching for battery cathode materials. *J. Chem. Phys.* **154**, 174111 (2021).
38. S. D. Cataldo, W. von der Linden, L. Boeri, First-principles search of hot superconductivity in la-xh ternary hydrides. *npj Comput. Mater.* **8**, 1–8 (2022).
39. F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, O. A. Von Lilienfeld, Prediction errors of molecular machine learning models lower than hybrid dft error. *J. Chem. Theory Comput.* **13**, 5255–5264 (2017).
40. K.-R. Müller, M. Finke, N. Murata, K. Schulten, S.-i. Amari, A numerical study on learning curves in stochastic multilayer feedforward networks. *Neural Comput.* **8**, 1085–1106 (1996).
41. F. Faber, A. Lindmaa, O. A. Von Lilienfeld, R. Armiento, Crystal structure representations for machine learning models of formation energies. *Int. J. Quantum Chem.* **115**, 1094–1101 (2015).
42. F. A. Faber, A. S. Christensen, B. Huang, O. A. Von Lilienfeld, Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* **148**, 241717 (2018).
43. M. Hellenbrandt, The inorganic crystal structure database (ICSD)–present and future. *Crystallogr. Rev.* **10**, 17–22 (2004).
44. S. Wang, Z. Wang, W. Setyawan, N. Mingo, S. Curtarolo, Assessing the thermoelectric properties of sintered compounds via high-through put *ab-initio* calculations. *Phys. Rev. X* **1**, 021012 (2011).
45. A. D. Sendek, Q. Yang, E. D. Cubuk, K.-A. N. Duerloo, Y. Cui, E. J. Reed, Holistic computational structure screening of more than 12 000 candidates for solid lithium-ion conductor materials. *Energ. Environ. Sci.* **10**, 306–320 (2017).
46. P. Canepa, G. S. Gautam, D. C. Hannah, R. Malik, M. Liu, K. G. Gallagher, K. A. Persson, G. Ceder, Odyssey of multivalent cathode materials: Open questions and future challenges. *Chem. Rev.* **117**, 4287–4341 (2017).

47. N. A. Zarkevich, D. D. Johnson, V. K. Pecharsky, High-throughput search for caloric materials: The caloricool approach. *J. Phys. D Appl. Phys.* **51**, 024002 (2017).
48. V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
49. M. I. Aroyo, J. M. Perez-Mato, C. Capillas, E. Kroumova, S. Ivantchev, G. Madariaga, A. Kirov, H. Wondratschek, Bilbao crystallographic server: I. Databases and crystallographic computing programs. *Z. Kristallogr. Cryst. Mater.* **221**, 15–27 (2006).
50. L. L. Boyle, J. E. Lawrenson, The origin dependence of wyckoff site description of a crystal structure. *Acta Crystallogr. A Cryst. Phys. Diffr. Theor. Gen. Crystallogr.* **29**, 353–357 (1973).
51. A. Togo, I. Tanaka, Spglib: A software library for crystal symmetry search. arXiv:1808.01590 [cond-mat.mtrl-sci] (5 August 2018).
52. D. Hicks, C. Oses, E. Gossett, G. Gomez, R. H. Taylor, C. Toher, M. J. Mehl, O. Levy, S. Curtarolo, AFLOW-SYM: Platform for the complete, automatic and self-consistent symmetry analysis of crystals. *Acta Crystallogr. A Found. Adv.* **74**, 184–203 (2018).
53. I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in *Proceedings of 7th International Conference On Learning Representations* (Openreview.net, 2019) -- Openreview.net is the publisher according to https://dblp.org/rec/conf/iclr/LoshchilovH19.html?view=bibtex.
54. D. A. Nix, A. S. Weigend, Estimating the mean and variance of the target probability distribution, in *Proceedings of 1994 IEEE International Conference On Neural Networks (ICNN'94)* (IEEE, 1994), vol. 1, pp. 55–60.
55. A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision?, in *Proceedings of Advances In Neural Information Processing Systems 30* (Curran Associates, Inc., 2017), pp. 5574–5584.
56. M. J. Mehl, D. Hicks, C. Toher, O. Levy, R. M. Hanson, G. Hart, S. Curtarolo, The aflow library of crystallographic prototypes: Part 1. *Comput. Mater. Sci.* **136**, S1-S828 (2017).
57. D. Hicks, M. J. Mehl, E. Gossett, C. Toher, O. Levy, R. M. Hanson, G. Hart, S. Curtarolo, The aflow library of crystallographic prototypes: Part 2. *Comput. Mater. Sci.* **161**, S1–S1011 (2019).
58. G. Kresse, J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15–50 (1996).
59. G. Kresse, J. Furthmüller, Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
60. P. E. Blöchl, Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
61. G. Kresse, D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
62. J. P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
63. A. Jain, G. Hautier, S. P. Ong, C. J. Moore, C. C. Fischer, K. A. Persson, G. Ceder, Formation enthalpies by mixing GGA and GGA + *U* calculations. *Phys. Rev. B* **84**, 045115 (2011).
64. R. Armiento, in *Database-Driven High-Throughput Calculations And Machine Learning Models For Materials Design* (Springer International Publishing, 2020), pp. 377–395.
65. S. P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder, K. A. Persson, The materials application programming interface (api): A simple, flexible and efficient api for materials data based on representational state transfer (rest) principles. *Comput. Mater. Sci.* **97**, 209–215 (2015).
66. S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, C. Wolverton, The open quantum materials database (oqmd): Assessing the accuracy of dft formation energies. *npj Computational Materials* **1**, 1–15 (2015).
67. V. Stevanović, S. Lany, X. Zhang, A. Zunger, Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies. *Physical Review B* **85**, 115104 (2012).
68. R. Friedrich, D. Usanmaz, C. Oses, A. Supka, M. Fornari, M. B. Nardelli, C. Toher, S. Curtarolo, Coordination corrected ab initio formation enthalpies. *npj Comput. Mater.* **5**, 1–12 (2019).
69. A. Wang, R. Kingsbury, M. M. Dermott, M. Horton, A. Jain, S. P. Ong, S. Dwaraknath, K. Persson, A framework for quantifying uncertainty in DFT energy corrections. *Sci. Rep.* **11**, 15496 (2021).
70. L. Breiman, Random forests. *Machine Learning* **45**, 5–32 (2001).
71. L. Ward, R. Liu, A. Krishna, V. I. Hegde, A. Agrawal, A. Choudhary, C. Wolverton, Including crystal structure attributes in machine learning models of formation energies via voronoi tessellations. *Phys. Rev. B* **96**, 024104 (2017).
72. P. B. Jørgensen, E. Garijo Del Río, M. N. Schmidt, K. W. Jacobsen, A. Agrawal, A. Choudhary, C. Wolverton, Materials property prediction using symmetry-labeled graphs as atomic position independent descriptors. *Phys. Rev. B* **100**, 104114 (2019).
73. Y. Zuo, M. Qin, C. Chen, W. Ye, X. Li, J. Luo, S. P. Ong, Accelerating materials discovery with bayesian optimization and graph deep learning. *Mater. Today*, (2021).

Acknowledgments: We thank J. Riebesell for valuable discussions regarding this work and H.-C. Wang and coauthors for providing the prerelaxation structures for the WBM dataset. **Funding:** R.E.A.G. and A.A.L. acknowledge the support of the Winton Programme for the Physics of Sustainability. A.A.L. acknowledges support from the Royal Society. F.A.F. acknowledges funding from the Swiss National Science Foundation (grant no. P2BSP2_191736). R.A. and A.S.P. acknowledge support from the Swedish Research Council (VR) grant no. 2020-05402 and the Swedish e-Science Centre (SeRC). The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at NSC partially funded by the Swedish Research Council through grant agreement no. 2018-05973. **Author contributions:** R.E.A.G.: Methodology, software, investigation, visualization, and writing—original draft. A.S.P.: Methodology, investigation, and validation. F.A.F.: Conceptualization, supervision, and writing—review and editing. R.A.: Conceptualization, supervision, and writing—review and editing. A.A.L.: Supervision and writing—review and editing. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** The relaxed and initial structures for the TAATA dataset are available at https://doi.org/10.5281/zenodo.5530535. The MP dataset used for this work was queried from https://materialsproject.org (*25*) using the Materials API (*65*). The data were queried from Supplemental Database Release V2021.03.22. The relaxed structures for WBM dataset used in this work are available at https://archive.materialscloud.org/record/2021.68. The prerelaxation structures were obtained from the original authors on request. The ICSD (*29*, *43*) can be accessed under license from FIZ Karlsruhe–Leibniz Institut for Information Infrastructure at https://icsd.products.fiz-karlsruhe.de/. The relaxed and initial structures for all completed prospective calculations are available at https://doi.org/10.5281/zenodo.6345276. An open-source implementation of the model is available at https://doi.org/10.5281/zenodo.6635011.

Submitted 24 November 2021
Accepted 13 June 2022
Published 27 July 2022
10.1126/sciadv.abn4117